Using Meta-data to search for Clinical Records: RMIT at TREC 2012 Medical Track

Iman Amini* Mark Sanderson* David Martinez† Xiaodong Li*
*RMIT Dept of Computer Science and NICTA, Australia
†NICTA and the University of Melbourne, CIS Department, Australia
iman.amini,mark.sanderson,xiaodong.li@rmit.edu.au
david.martinez@nicta.edu.au

Abstract

Clinical records contain International Classification of Diseases (ICD) codes summarizing the primary and/or secondary diseases of patients; these codes can be used as evidence to find relevant documents. In this paper we propose a novel approach to locally build a knowledge source from the existing data set to be used for query expansion, exploiting the ICD codes. While this approach does not rely on any external knowledge sources, it proves to be significantly superior to our global expansion and baseline systems.

1 Introduction

The vocabulary gap between the language of documents and queries is one of the facts which deteriorates the effectiveness of a retrieval system. This can be more notable for retrieval in the clinical domain where practitioners often use abbreviations and synonyms to refer to medical concepts.

Automatic query expansion is one of the most prevalent approaches in Information Retrieval (IR) to enhance the effectiveness of retrieval systems and address the problem of vocabulary gap. The existing expansion methods can be categorized as either local or global, where each refer to the source where extra information can be extracted from. The local expansion methods exploit terms only from the underlying collection, while global methods can extract information from external knowledge sources as well. While both approaches can be effective in improving a subset of the topics, they almost invariably exacerbate a subset of original topics as well. Most of the existing approaches apply a unified method to any given topic, albeit, some topics may require a different way of handling, while others may not even need expansion.

Finding a reliable way to introduce new terms to the queries can be quite challenging. However, clinical records in general may contain a rich set of meta-data, such as ICD codes which are assigned by practitioners to describe and/or summarize the primary or secondary diseases of the patients. This concise piece of information can be a reliable evidence in the process of finding a relevant clinical record and therefore a sensible way to add terms from, to the original queries. Nonetheless, these codes must be correctly inferred from the queries before finding documents that would contain the same.

Assigning correct ICD codes to queries can also be a difficult task. Queries can describe several aspects, and ICD codes have a hierarchical structure with different levels of specificity, making it hard to automatically assign the most correct ICD code to each aspect of the query. For instance *hearing loss* as an aspect of a query can be linked to many ICD codes, including but not limited to 389.03, 389.0, 380.01, where each code is concerned with a different type of hearing problem. The three codes in the example refer to, middle ear, conductive and external hearing loss respectively. On the other hand for some more specific aspects of the queries, there may be no more than one ICD code.

The main contribution of this paper is introducing a new approach which suggests a possible way to map the plain queries into ICD codes and then use the same to find relevant documents. These relevant documents will be used as a local source to selectively assign informative terms and enrich the plain queries. This technique is in a way analogous to the popular pseudo relevance feedback where a local set of elite documents are chosen to extract informative terms. Secondly we perform an analysis on the data set, focusing on two aspects of the search for clinical records: query expansion and negation handling. These analyses were aimed to answer two research questions:

- Is query expansion necessary for an ideal clinical system to use for Medical TREC?
- Is negation handling vital for a clinical search for the data set used in the Medical TREC?

To answer the first question we considered the fact that, if queries and documents share a very similar vocabulary and the terms from the external sources do not largely occur in the collection, we can infer that query expansion can be almost rendered ineffective. However, we were aware about the notable diversity in concepts and terms used by the medical community to describe the same concept. Nonetheless, we intended to investigate whether this large variety of terms are used in the collection.

To find an answer to our second question we looked at the distribution of negated query terms in the relevant and

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
NOV 2012

2. REPORT TYPE

3. DATES COVERED
00-00-2012 to 00-00-2012

4. TITLE AND SUBTITLE
Using Meta-data to search for Clinical Records: RMIT at TREC 2012 Medical Track

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
RMIT University,Dept of Computer Science and NICTA,Melbourne VIC 3001 Australia,

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT
Clinical records contain International Classification of Diseases (ICD) codes summarizing the primary and/or secondary diseases of patients; these codes can be used as evidence to find relevant documents. In this paper we propose a novel approach to locally build a knowledge source from the existing data set to be used for query expansion, exploiting the ICD codes. While this approach does not rely on any external knowledge sources, it proves to be significantly superior to our global expansion and baseline systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**9**	

non-relevant documents per topic. This analysis answered the question of whether negation can cause a higher recall or will it just eschew the retrieval of non-relevant documents. Although Natural Language Processing (NLP) approaches such as negation detection are considered vital in the clinical domain of IR, they have been shown to be largely unsuccessful for the general domain in IR [20].

Our runs this year revolve around the use of meta data, ICD codes in particular, to enrich the otherwise plain queries. This idea was formed when we needed some form of pseudo relevance judgement to evaluate our systems while the sample assessments were not enough in the first year of Medical TREC in 2011. The positive correlation of our pseudo and official relevance judgment encouraged us to use the artificially built set of judgments, to build a method similar to the pseudo relevance feedback.

We structured the paper in the following way, analysis of the data is explored and relevant statistics are reported first, next we explain our methodologies for query expansion using external knowledge sources and ICD codes and results are provided in the subsequent sections. Finally we conclude the paper with the final rankings of Medical TREC from NIST.

2 Background

This section provides a brief survey of the approaches taken by the top scoring teams in TREC 2011 [10, 14] who demonstrated that significant improvements are dependent on smart expansion and NLP techniques to effectively enrich the queries and exploit the structure of the medical reports. Cengage learning [14] submitted the best automatic run, suggesting that information extraction and query expansion hold the key to a reliable retrieval system. However, our analysis highlights the old fact about the risk of over expansion for some topics while suggesting an approach for preventing this problem.

King et al. [14] developed an extensive pipeline to preprocess text and apply NLP techniques such as lemmatization and part of speech tagging, while using UMLS[1] to label and relate terms for expansion. Another factor that contributed to the effectiveness of their system was creating their own knowledge source, Cengage collection of medical reference encyclopedia, which was eventually used to resemble traditional pseudo relevance feedback. Also in their work, the problem of synonymy was tackled by indexing UMLS Concept IDs, which would match all the descriptions that map to the same concept ID.

A comprehensive semantic search was carried out by Gurulingappa et al. [10], resulting in very competitive scores. They found that relying on the Metamap [3] software for concept identification was counterproductive, indicating the existence of false positives. This problem was addressed by using ProMiner [11]. Their manual assessments of top 10 retrieved documents by some of their systems revealed some errors in TREC's manual judgements.

Goodwin et al. [9] focused on query analysis and reformulation to extract age and gender related terms. To bridge the vocabulary gap between queries and documents they select keywords from the given topics using Wikipedia and then expand them with Pubmed, UMLS and SNOMED as external sources. While most groups including RMIT [2] performed sanity checks at the preretrieval stage, they take an extra precautionary step of filtering out retrieved documents that contradict the inclusion criteria, using age, gender and negation detection algorithms. For expansion they manually assign weights to terms from different Knowledge Sources (KSs), whereas Zhu and Carterette [21] use an algorithm known as CORI [6] to automatically assign weights to external KSs, by means of assigning similarity weights based on term overlap. In their work three sources that were not used by other teams, to the best of our knowledge, were tested. According to the TREC official results, Genomic TREC and one day PubMed query log were most and least effective respectively, while image CLEF had moderate improvement over their baseline system.

Schuemie et al. [19], after their analysis of the collection, found that discharge summary sections are not always reliable and instead they found the *postoperative diagnosis* section to be more informative. Their best run was based on balancing the weights between the terms expanded from Wikipedia, using the Match Score Maximization algorithm which was designed to make certain that top results contain most aspects of the queries. This was inspired by the conclusion made by the Reliable Information Access (RIA) Workshop [12] where a task of failure analysis per topic basis was explored. After exhaustively analyzing 45 topics, 10 categories to describe the limitation of the IR systems were drawn and Buckley [4] concludes that all categories share the same reason, which is the failure to retrieve documents that contain all aspects of the queries.

The University of Glasgow [17], experiments with a novel voting model approach, and propose a simple way to implement negation handling by adding a prefix to the negated concepts in the reports, while RMIT [2] and NICTA [13] implemented a similar approach, others mostly eliminated negated concepts from the reports. Additionally the weight of the expanded terms from Wikipedia and MeSH, were calculated using EMIM (Expected Mutual Information Measure). Their best run maps the ICD codes in the visits and expands the available concepts with Wikipedia, which was a unique way to expand documents with a knowledge source rather than queries.

3 Analysis over the TREC 2011 task

In this section we report the result of our analysis on two aspects of a clinical search mentioned earlier in this paper. A list of all the terms and phrases that we extracted from the 35 test topics and used for the sake of our analysis is presented in the last page of this paper. We used the test collection from TREC 2011 throughout this section.

[1] http://www.nlm.nih.gov/research/umls/

Query Number	Relevant	Non-relevant
101	0	26
102	0	18
103	0	141
104	0	9
105	0	61
106	0	251
107	0	2
108	0	573
109	0	21
110	0	0
. . .	. . .	. . .
Mean	0	195.8

Table 1: Number of negative medical terms found in each relevant and non-relevant set for 35 topics

3.1 Negation

In this section we show the figures collected by counting the number of negated occurrences of the query terms in the relevant and non-relevant sets. We wanted to explore how important is to use negation while earlier we found a rational way to implement it [2].

Identifying negated concepts from the documents or queries have been proved vital for medical records in the past research [14, 15]. Negex[8] is a rule based algorithm which is largely used for identification of negated concepts and is reported to be reliable in finding true positives and obtaining high scores on sensitivity and precision when finding negative phrases in the clinical data [7] and due to its popularity we utilize it to automatically find negated concepts in the collection.

In general, there are two ways to handle negated concepts, either discarding them from the queries before retrieval or detecting them from the documents prior or while indexing. We settle for using the pre-indexing approach and use a rule based approach to substitute the negated terms with a new prefix. In order to convert the original queries to a set of keywords for the sake of counting, we use Metamap again to decide which query terms to keep.

The numbers in the Table 1 show that NegEx can be mainly effective for filtering out non-relevant documents rather than assisting to find more relevant documents for a given topic. Relevant documents tend to be devoid of negated concepts found in the queries, and this can be expected. However, the 0 encounter of negated terms across the relevant set, and the fact that NegEx was clearly separating out the relevant from non-relevant set was surprising and forced us to double check the counts for confirmation. From this table we can clearly answer **No** to the question, *is a document with negated query term relevant?* Nonetheless, the impact of handling negated terms with regards to the overall performance of retrieval systems may depend on the type of collection and the degree of used negated terms.

Due to space constraints only first 10 queries are shown with their corresponding mean for the whole 35 test queries. Relevant documents are those found from the relevance judgement and all other documents in the collection are considered to be non-relevant.

We used the keywords to count the number of occurrence using our Metamap identified concepts from the topics. This set is provided in the last page of the paper. The count for each topic is the summation of the counts of all the query terms for that particular topic. Note that Metamap is a software which we use frequently in this paper to identify and disambiguate medical concepts.

3.2 Query Expansion with external knowledge sources

In this section we first explain the process by which we find terms for expansion from UMLS and Dbpedia, next we state how we refine and combine these sources to build a system for global expansion.

3.2.1 Finding expansion terms from Dbpedia and UMLS

The first step involved finding the terms that needed expansion, in other words, we stopped terms that did not fall under any medical categorical concepts, and we call these concepts as candidate terms. We used *MetaMap-2010* to identify phrases linked to terms in the UMLS Metathesaurus (version 2010AA), and we only keep terms that have a mapping to a medical concept. For example, for the query below we kept these phrases: "liver metastasis" "treated hospital" "cancer patient" "procedure", and this is more effective than just using stop words, as we can identify phrases from the queries that exist in the UMLS data source.

Cancer patients with liver metastasis treated in the hospital who underwent a procedure

In some cases terms consisted of a primary term followed by a parenthesized description — such as "Intervention (Surgical and medical procedures)" — and in such cases we treated them as separate candidate terms. To this end we have a bunch of phrases and terms ready to be expanded. In the case of DBpedia[2], we extracted all the terms listed under the "category" for each phrase, provided that a matching web page existed for the given term. We treated the term to be expanded as a category, and retrieved all the terms that fall under it. Finally we removed terms with the following strings from the DBpedia output: "code", "history", "mechanism", "poisoning", "toxicity", and "withdrawal".

During the development process, we also explored expansion using hierarchical relations from the UMLS Metathesaurus, by selecting all the terms in the hyponym concepts. A hyponym of a word has a more specific meaning than the word associated to it, and we found that some medical terms in the queries can have up to 4000 hyponyms. However, we observed that DBpedia offered a higher coverage of some domains, such as newly developed drugs, and less risk of over-expansion. For instance, one sample query contained the term "atypical antipshychotic", which UMLS expanded with 8 more specific drugs (e.g. "Clozapine"). DBpedia, however, identified the same set of drugs as well as a further 22 new drug and brand names, which seemed correct after manual analysis, and had a stronger presence in the collection.

[2] http://wiki.dbpedia.org/OnlineAccess

Knowledge Source (KS)	Count(no-stem)	Count(stemmed)	Bpref(stemmed)	Bpref(combined with original query terms)
Original Query Terms (without KS)	56.83	59.52	0.3143	N/A
Dbpedia	42.61	55.95	0.2831	0.3048
Wikipedia	23.11	23.48	0.0903	0.2866
UMLS (Hyponyms)	55.51	56.56	0.2044	0.2874

Table 2: Percentage and mean Bpref of exact matches using query terms and extended terms from different knowledge sources across the relevant documents

3.2.2 Choosing the right knowledge source for Query Expansion

Expanding queries is an intuitive attempt to increase the chance of retrieving extra relevant documents, and a good approach should dampen the risk of worsening the original query. While there exist a plethora of resources to use for global expansion, it is crucial to find the most reliable external knowledge source to extract relevant query terms. If the source for query expansion is external and the language between queries and the knowledge source do not match sufficiently, then expanding queries is most likely to deteriorate the search outcome than to improve it. In this section we measure the degree of overlap between the language of the query, before and after expansion and the collection. Our aim was to determine firstly, which of the several knowledge sources have the highest degree of overlap with the language of the documents and secondly, how much does the overlap correlate with the performance of the search.

We compare three external sources, Wikipedia, Dbpedia and hyponyms from the UMLS and count the number of exact matches of the expanded and original terms in the relevant documents, for each query. In order to verify if there is a correlation between the vocabulary overlap and the performance of the search we calculate the Bpref, first by using terms only from the queries and knowledge sources, and then by combining original terms to the expanded terms. The original terms used for this experiment are the list of keywords that we gathered from Metamap which is made available as appendix to this paper.

We show in Table 2 the coverage of query terms across the relevant documents before and after we use different types of expansion and stemming. The numbers indicate that the vocabulary that has generated the queries is more likely to be the same as the documents, comparing to other knowledge sources. This is followed by the hyponyms extracted from the UMLS. However, looking at the performances given in the last two columns, we can see that all the knowledge sources, when used on their own right and combined with the original query terms, deteriorate the performance. This implies the presence of noise introduced from all the sources which negatively affects the scores. This is highly noticeable for the case of Wikipedia and therefore we decided to discard this knowledge source and attempt to minimize the noise from other sources by using semantic types, which is explained in the next section.

Semantic Types (STs)	**Dbpedia**	**UMLS(hyponym)**
Disease or Syndrome	0.3371	**0.3145**
Finding	0.304	0.3111
Functional Concept	0.3061	0.3077
Medical Device	**0.3374**	-
Spatial Concept	-	0.3134
Activity	-	-
Baseline(No Expansion)	0.3143	0.3143

Table 3: Rank and Score using different semantic selection to expand queries

3.3 Using UMLS semantic types to filter unwanted terms for query expansion

Apart from the difficulty in finding a reliable source for expansion, there lies the problem of selecting the most informative terms. While pseudo relevance feedback exploits the assigned weights to the terms in the document and effectively select only the top n terms, words extracted externally are not weighted, making it hard to discard the potential noise.

Semantic Types (STs) are assigned to medical concepts in the UMLS relational tables and they are extractable from the output generated by the Metamap. The idea to use STs was driven by the problem of extracting extremely long queries when using knowledge sources such as Wikipedia or UMLS for query expansion. Some terms, when expanded, resulted in a large bundle of words which was not a practical size for a typical query. Therefore it was essential to find an optimal way to reduce the length of expansion while preserving the most informative terms. We wanted to find the most informative STs that point to useful terms to keep, while filtering out the rest.

NICTA [13] was the other TREC participant who took a similar approach by intuitively choosing two STs and discarding the remaining expanded terms with other types. Last year we [2] drew 61 STs out of all the 35 queries and ran an experiment to individually get the score of each ST in order to systematically select the best scoring STs. However, we filtered out terms before expansion, meaning that we only expanded terms that had a particular ST, on the contrary NICTA chose acceptable terms from the already expanded terms.

Table 3 shows the top 4 semantic types for UMLS and Dbpedia that provide the best results after being added to the plain queries. All the scores are given in Bpref and the highest results are differentiated in bold. For instance, *Disease or Syndrome* ST is the best for UMLS, while it is the second best for terms extracted from the Dbpedia. We only show the top four scoring STs in this table.

The low performance obtained by using some of the STs, those for UMLS in particular, were tracked down to large size of the expansion for only a fraction of the topics. This

V58.66	Long-term (current) use of aspirin
596.7	Hemorrhage into bladder wall
585.6	End stage renal disease
786.8	Hiccough
941.13	Erythema due to burn (first degree) of lip(s)
783.40	Unspecified lack of normal physiological development
V44.4	Status of other artificial opening of gastrointestinal tract
952.08	C5-c7 level with central cord syndrome

Table 4: ICD codes and descriptions

implies that we still need a more precise approach for refinement of such cases. However, for our participation in TREC 2012 we decided to focus on a safer technique for expansion which is explained in section 3.4

For our final runs we discarded all the terms associated with *Functional Concept* firstly because it led to over-expansion and secondly the terms from this category described very general aspects of the topics which did not need expansion. The pipeline for combining the expansion types is demonstrated in Figure 1. A comparison of this expansion and our baseline is given in section 4.

3.4 Pseudo Relevance Feedback with ICD codes

The Medical records provided in the TREC collection embody two tags with ICD codes, relevant to discharge and admit diagnosis of the patients. These codes can be a good summary of the whole document when mapped into their corresponding descriptions. ICD map tables with full descriptions are publicly available and can be used to expand the ICD codes interspersed in the medical reports. An example of an ICD table can be seen in Table 4. While these codes can well summarize a particular record, they lack information about the population, age and details of patients. Therefore they can't be used as the only evidence for finding the most relevant documents.

In order to automatically pick the best ICD codes for each query we treated the description of ICD codes as documents and index them using the Terrier search engine, where document ids are taken to be the ICD codes. The next step is retrieval and 35 test queries from TREC 2011 are used to retrieve ICD codes. A ranked list of documents are returned from which the most likely ICD codes are chosen, however, choosing the number of ICD codes was done empirically and we optimized the length of ICD codes by extrinsically evaluating it using our training data. To this end we have some evidence to find relevant documents, and we use regular expression to gather all the reports that have at least one of the top-n ICD codes. We had to use the logical OR for n > 1, as further restriction caused retrieving a null set for most of the topics. Finally we use the Bose-Einstein-1 [1] weighting model from the Divergence From Randomness (DFR) framework and select 40 terms from these elite documents in our pseudo relevance feedback. All the terms will be given weights calculated by BE1 weighting model. The steps from ICD code extraction until building a pseudo relevance feedback are shown in Figure 2.

Length	Bpref	Map	P_10
1	**0.5450**	**3775**	0.5171
2	0.5231	0.3644	0.5057
3	0.5212	0.3662	0.5171
4	0.4937	0.3456	**0.5400**
5	0.4895	0.3494	0.5371

Table 5: Choosing the best length in the ICD ranked list

3.5 Choosing the best number of ICD codes for finding relevant documents

Medical records can contain 0 to 15 ICD codes [16] and the order of their occurrence is mostly insignificant. Nonetheless the number of ICD codes to describe an ad-hoc query does not exceed 3 to 4 codes. After we automatically find the best ICD codes for each topic, however, we need to decide for what value of n we choose the top-n ICD codes to build our pseudo relevance feedback.

We experiment with 5 varying lengths of ICD codes, that is, for each topic we assign top n ICD codes retrieved by the search engine, for n ranging from 1 to 5. Table 5 shows the effect of the varying number of ICD codes assigned to each query for selecting the relevant documents. It can be seen that increasing the number of ICD codes as evidence to find the relevant documents is counterproductive. However, this trend seems to be the opposite, in case of precision at 10 and the longer the length of ICD codes the higher the score, with the exception for length 2 and 5. Note that we use a logical OR rather than AND to select documents that contain at least one of the given ICD codes where the number of ICD codes is greater than 1. Using AND proved far too exclusive and failed to find documents for more than 60% of the topics.

4 Analysis of RMIT official runs

TREC received 88 runs this year, all of which contributed to the judgment sets, while a more effective sampling of runs was performed this year which allowed NIST to report inferred measures.

Two of RMIT runs, APRel1 and APRel2 performed above the mean, and the ordering of the official scores are mostly consistent with our results using the training data from TREC 2011. Our best run, APRel1, was found to be statistically significant at the 0.05 level as compared to GE4 and our baseline.

Table 6 is a summary of configuration and Bpref scores gathered using the training data from TREC 2011 and Table 7 is the official scores from NIST for TREC 2012. Our baseline system has the same configuration as other techniques, except that it is not expanded, rather it is all the plain query terms with the non-informative terms stopped using the Metamap software. For indexing and retrieval we relied on the Terrier [18] open source search engine and the ranking algorithm was fixed to be PL2 which is one of the models from the Divergence from Random framework.

Our first three runs (APRel1,APRel2,RARP2) were based on ICD expansion in conjunction with pseudo relevance feed-

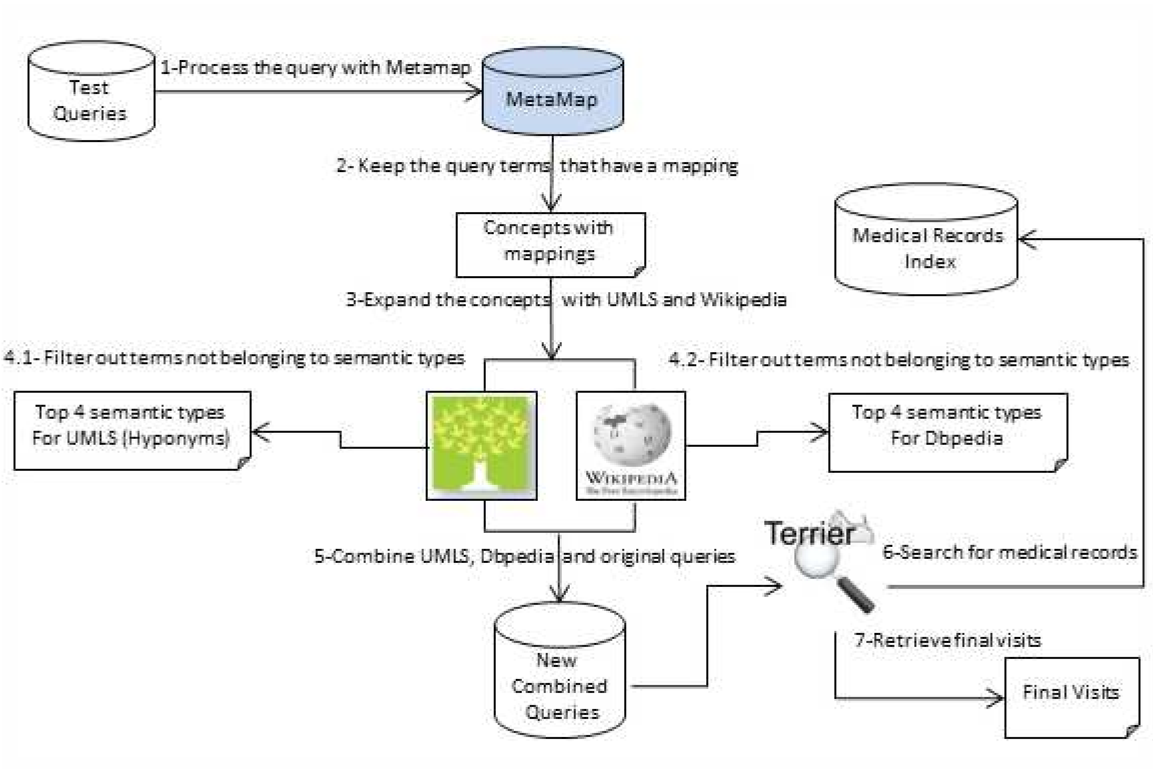
Test Queries
1-Process the query with Metamap
MetaMap
2- Keep the query terms that have a mapping
Concepts with mappings
Medical Records Index
3-Expand the concepts with UMLS and Wikipedia
4.1- Filter out terms not belonging to semantic types
4.2- Filter out terms not belonging to semantic types
Top 4 semantic types For UMLS (Hyponyms)
Wikipedia
Top 4 semantic types For Dbpedia
Terrier
5-Combine UMLS, Dbpedia and original queries
6-Search for medical records
7-Retrieve final visits
New Combined Queries
Final Visits

Figure 1: Combining knowledge sources

	Submitted Runs				
	APRel1	APRel2	RAPRel2(Ranked_APrel2)	GE4	Baseline
Expansion	✓	✓	✓		
ICD-Length	1	2	2	N/A	N/A
Stopped	✓	✓	✓	✓	✓
Negation-Aggressive	✓	✓	✓	✓	✓
Ranking model	PL2	PL2	PL2	PL2	PL2
Indexing	R	R	R	R	R
Stemming	Porter	Porter	Porter	Porter	Porter
Bprefs in Training Data	0.5450	0.5231	0.5193	0.3457	0.3143

Table 6: Specifications of RMIT's four official runs In *indexing* R represents report-based indexing.

back. They are mainly different in the number of ICD codes used to gather relevant documents to build a pseudo relevance feedback for each query. RAPRel2, is also different with the first two runs, as it also ranks the documents in pseudo relevance set based on the number of relevant ICD codes and the overall length of codes in the documents, selecting only the top 10 documents.

Using only the top ranked ICD code for each query consistently proved most effective followed by using two. The rationale behind it is that, we used a logical OR to gather documents that contained at least one of the ICD codes, this led to the creation of an enlarged pseudo relevance feedback source which introduced extra noise from the second ICD code which may not have been the most relevant to the query. RAPRel2 was built to address this problem, ranking the retrieved documents in the pseudo relevance set by dividing the number of relevant ICD codes by the overall length of ICD codes in the documents. However we did not optimize the desired number of documents for each query to effectively build pseudo relevance feedback. Rather we chose the top 10 documents for each query and this needs further work to determine the most effective number of documents to get the best possible outcome. We can potentially learn that by looking at the number of documents used for the APRel1 run. The fourth run which does not incorporate any meta-data from the documents, is a combination of expanded terms from external knowledge sources which was explained in section 3.3. Although we automatically chose 4 semantic types to help with filtering out non informative terms and to reduce the length of expansion, some concepts were linked with a large number of relevant terms. We manually looked at the expanded queries and observed that the low performance of most of the topics were linked to the large size of expansion.

Metric	**APRel1**	**APRel2**	**RAPRel2**	**GE4**	**Mean**
InfAP	0.1757	0.1711	0.1405	0.1429	0.1689
Bpref	0.2855	0.2812	0.2556	0.2447	-

Table 7: Evaluation of RMIT's 4 official runs

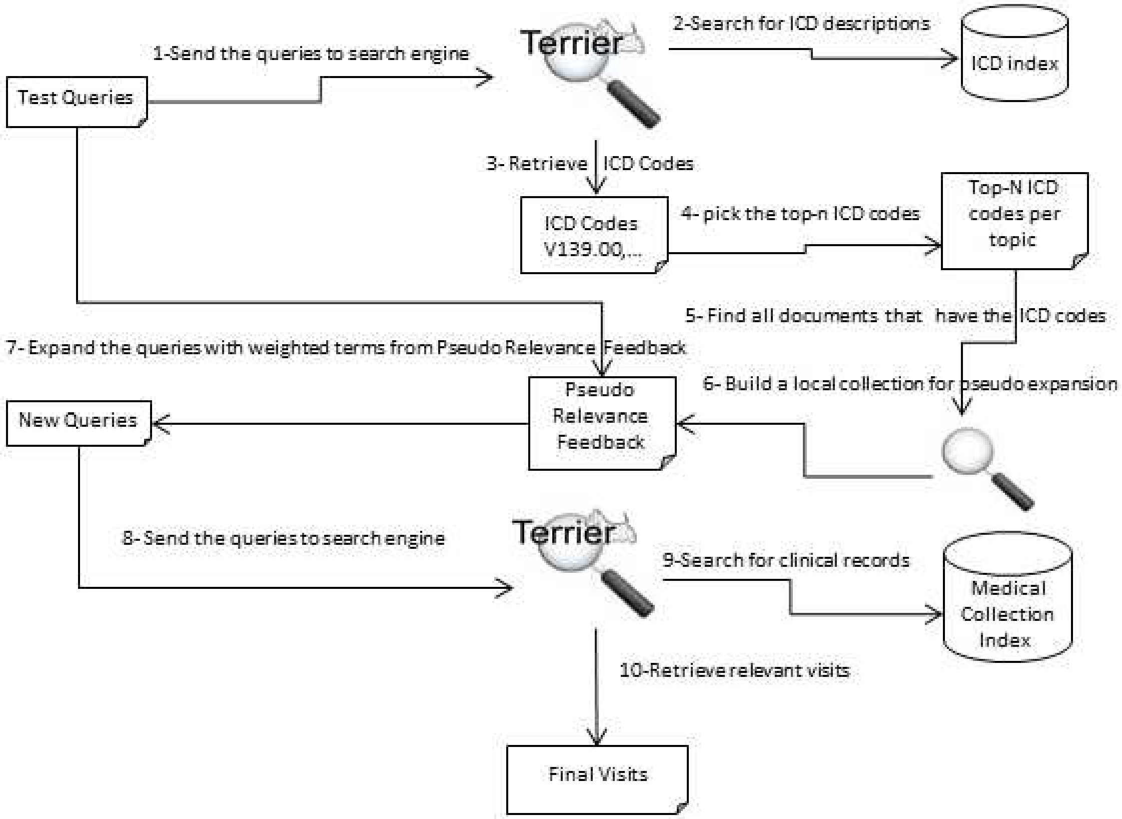
Test Queries
1-Send the queries to search engine
Terrier
2-Search for ICD descriptions
ICD index
3- Retrieve ICD Codes
ICD Codes V139.00,...
4- pick the top-n ICD codes
Top-N ICD codes per topic
5- Find all documents that have the ICD codes
7- Expand the queries with weighted terms from Pseudo Relevance Feedback
6- Build a local collection for pseudo expansion
New Queries
Pseudo Relevance Feedback
8- Send the queries to search engine
Terrier
9-Search for clinical records
Medical Collection Index
10-Retrieve relevant visits
Final Visits

Figure 2: ICD extraction and pseudo relevance feedback life-cycle

5 Conclusion

We developed and tested a new approach for locally expanding queries using ICD codes. This simple technique was found to be statistically significant over our baseline and global expansion method, suggesting the usefulness of incorporating the existing meta-data from the documents rather than relying merely on the external knowledge sources for query expansion.

We learned about the reasons behind the ineffectiveness of our external expansions, which was tracked down to the extreme size of expansion and the degree of noise introduced for some particular topics. In our future work we will focus on finding a reliable way to weight all the terms extracted from each external sources and to effectively limit the size of expansion.

We also intend to extend our work to improve the ICD expansion by using extra meta-data to compensate for the complicated aspects of the queries which do not have any corresponding ICD codes. While we need to find a more reliable way for mapping queries into ICD codes to make sure that all aspects in the queries map to the most relevant ICD code.

6 Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] G. Amati. *Probability models for information retrieval based on divergence from randomness.* PhD thesis, University of Glasgow, 2003.

[2] I. Amini, M. Sanderson, D. Martinez, and X. Li. Search for Clinical Records: RMIT at TREC 2011 Medical Track. In *Proceedings of Text Retrieval Conference*, 2011.

[3] A.R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[4] C. Buckley. Why current ir engines fail. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 584–585. ACM, 2004.

[5] C. Buckley and E.M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32. ACM, 2004.

[6] J. Callan. Distributed information retrieval. *Advances in information retrieval*, pages 127–150, 2002.

[7] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium*, page 105. American Medical Informatics Association, 2001.

[8] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[9] T. Goodwin, B. Rink, K. Roberts, and S.M. Harabagiu. Cohort Shepherd: Discovering Cohort Traits from Hospital Visits. In *Proceedings of TREC*, 2011.

[10] H. Gurulingappa, B. Müller, M. Hofmann-Apitius, and J. Fluck. A Semantic Platform for Information Retrieval from E-Health Records. In *Proceedings of TREC*, 2011.

[11] D. Hanisch, K. Fundel, H.T. Mevissen, R. Zimmer, and J. Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(Suppl 1):S14, 2005.

[12] D. Harman and C. Buckley. The nrrc reliable information access (ria) workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 528–529. ACM, 2004.

[13] S. Karimi, D. Martinez, S. Ghodke, L. Zhang, H. Suominen, and L. Cavedon. Search for Medical Records: NICTA at TREC 2011 Medical Track. In *Proceedings of TREC*, 2011.

[14] B. King, L. Wang, I. Provalov, and J. Zhou. Cengage Learning at TREC 2011 Medical Track. In *Proceedings of TREC*, 2011.

[15] B. Koopman, P.D. Bruza, L. Sitbon, and M.J. Lawley. Analysis of the effect of negation on information retrieval of medical data. In *Proceedings of 15th Australasian Document Computing Symposium (ADCS)*. University of Melbourne, 2010.

[16] L.S. Larkey and W.B. Croft. Automatic assignment of icd9 codes to discharge summaries. *Center for Intelligent Information Retrieval Technical Report*, 1995.

[17] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M.M. Bouamrane. University of glasgow at medical records track: Experiments with terrier. In *Proceedings of TREC*, 2011.

[18] C. Macdonald, R. McCreadie, R.L.T. Santos, and I. Ounis. From puppy to maturity: Experiences in developing terrier. *Open Source Information Retrieval*, page 60, 2012.

[19] M. Schuemie, D. Trieschnigg, and E. Meij. Dutchhattrick: semantic query modeling, context, section detection, and match score maximization, 2011.

[20] E. Voorhees. Natural language processing and information retrieval. *Information Extraction*, pages 724–724, 1999.

[21] D. Zhu and B. Carterette. Using Multiple External Collections for Query Expansion. In *Proceedings of TREC*, 2011.

”hearing loss” patients
complicated receive gerd endoscopy patients
treated endocarditis ”methicillin-resistant staphylococcus aureus” ”hospitalized patients” mrsa
”localized cancer” robots diagnosed prostate patients ”prostate cancer” treated surgery localised
dementia patients
staging ct monitoring pet patients cancer ”positron-emission tomography” ”computed tomography” ”magnetic resonance imaging”
”ductal carcinoma” dcis patients
treated vascular surgically claudication patients
women osteopaenia
discharge hospital haemodialysis patients
”chronic back pain” patients intraspinal pump receive ”pain medicine”
female mastectomies admission ”breast cancer” patients
patients received adenocarcinoma admission revealed adult colonoscopies
patients hospice discharge home adult ”palliative care” admit adult asthma patients
hospital methotrexate received ”cancer treatment” patients
”post-traumatic stress disorder” patients
admission received ”stent, coronary” adults
”anion gap acidosis” patients ”emergency room” ”insulin dependent diabetes” secondary presented adult
chf treatment admit patients
patients give ”emergency department” ”acute coronary syndrome” presented cad plavix
”parenteral nutrition, total” hospital received patients
received hospital diabetic ”diabetic education” patients
episodes hospital patients present acute glaucoma secondary vision loss
infected hiv ”hepatitis c” patients
diagnosis admit ”multiple sclerosis” patients
”obesity, morbid” diabetes ”secondary disease” hypertension patients admit
patients medications ”knee surgery nos” post ”coagulant, nos” hip admit treated
”ct angiography” admit ”chest pain” assessed patients
”physical therapy” children received ”cerebral palsy” admit
”abdominal surgery” invasive patients
fusion patients discectomy admit surgery ”cervical spine”
take patients herbals osteoarthritis products admit care
patients ”chronic disorder” ”seizure disorder” seizure ”seizure activity” chronic admit control
”liver metastasis” treated hospital ”cancer patient” procedure

Table : List of all the words and phrases identified by Metamap and used for all the statistical analysis